# Convolutional Scale Invariance for Semantic Segmentation

Ivan Krešo[✉], Denis Čaušević, Josip Krapac, and Siniša Šegvić

Faculty of Electrical Engineering and Computing,
University of Zagreb, Zagreb, Croatia
ivan.kreso@fer.hr

**Abstract.** We propose an effective technique to address large scale variation in images taken from a moving car by cross-breeding deep learning with stereo reconstruction. Our main contribution is a novel scale selection layer which extracts convolutional features at the scale which matches the corresponding reconstructed depth. The recovered scale-invariant representation disentangles appearance from scale and frees the pixel-level classifier from the need to learn the laws of the perspective. This results in improved segmentation results due to more efficient exploitation of representation capacity and training data. We perform experiments on two challenging stereoscopic datasets (KITTI and Cityscapes) and report competitive class-level IoU performance.

## 1 Introduction

Semantic segmentation is an exciting computer vision task with many potential applications in robotics, intelligent transportation systems and image retrieval. Its goal is to associate each pixel with a high-level label such as the sky, a tree or a person. Most successful approaches in the field rely on dense strongly supervised multi-class classification of the image window centered at the considered pixel [28]. Such pixel-level classification of image windows resembles the localization task [30], which is also concerned with finding objects in images. However, the two tasks are trained in a different manner. Positive object localization windows have a well-defined spatial extent: they are tightly aligned around particular instances of the considered class. On the other hand, the class of a semantic segmentation window is exclusively determined by the kind of the object which projects to the central pixel of the window. Thus we see that the window size does not affect semantic segmentation outcome, which poses contrasting requirements. In cases of featureless or ambiguous texture, large windows have to be considered in order to squeeze information from the context [7]. In cases of small distinctive objects one has to focus onto a small neighborhood, since off-object pixels may provide misleading classification cues. This suggests that a pixel-level classifier is likely to perform better if supplied with a local image representation at multiple scales [4,11,21] and multiple levels of detail [22,24].

We especially consider applications in intelligent transportation systems and robotics. We note that images acquired from vehicles and robots are quite different from images taken by humans. Images taken by humans always have a purpose:

© Springer International Publishing AG 2016
B. Rosenhahn and B. Andres (Eds.): GCPR 2016, LNCS 9796, pp. 64–75, 2016.
DOI: 10.1007/978-3-319-45886-1_6

the photographer wants something to be seen in the image. On the other hand, a vehicle-mounted camera operates independently from the pose of the objects in the scene: it simply acquires a fresh image each 40 milliseconds. Hence, the role of the context [7] in car-borne datasets [6,12] will be different than in datasets acquired by humans [9]. In particular, objects in car-borne datasets (cars, pedestrians, riders, etc.) are likely to be represented at a variety of scales due to forward camera motion. Not so in datasets taken by humans, where a majority of objects is found at particular scales determined by rules of artistic composition. This suggests that paying a special attention to object scale in car-borne imagery may bring a considerable performance gain.

One approach to address scale-related problems would be to perform a joint dense recovery of depth and semantic information [8,20]. If the depth recovery is successful, the classification network gets an opportunity to leverage that information and improve performance. However, these methods have limited accuracy and require training with depth groundtruth. Another approach would be to couple semantic segmentation with reconstructed [23] or measured [2] 3D information. However, the pixel-level classifier may not be able to successfully exploit this information. Yet another approach would be to use the depth for presenting better object proposals [2,5]. However, proposing instance locations in crowded scenes may be a harder task than classifying pixels.

In this paper we present a novel technique for scale-invariant training and inference in stereoscopic semantic segmentation. Unlike previous approaches, we address the scale-invariance directly, by leveraging the reconstructed depth information [13,32] to disentangle the object appearance from the object scale. We realize this idea by introducing a novel scale selection layer into a deep network operating on the source image pyramid. The resulting scale-invariance substantially improves the segmentation performance with respect to the baseline.

## 2 Related Work

Early semantic segmentation work was based on multi-scale hand-crafted filter banks [26,28] with limited receptive fields (typically less than $50 \times 50$ pixels). Recent approaches [3,4,21,22] leverage amazing power of GPU [11] and extraordinary capacity of deep convolutional architectures [19] to process pixel neighborhoods by ImageNet-grade classifiers. These classifiers typically possess millions of trainable parameters, while their receptive fields may exceed $200 \times 200$ pixels [29]. The capacity of these architectures is dimensioned for associating an unknown input image with one of 1000 diverse classes, while they typically see around million images during ImageNet pre-training plus billions of patches during semantic segmentation training. An architecture trained for ImageNet classification can be transformed into a fully convolutional form by converting fully-connected layers into equivalent convolutional layers with the same weights [15,22,27]. The resulting fully convolutional network outputs a dense $\mathrm{W/s} \times \mathrm{H/s} \times 1000$ multi-class heat map tensor, where $\mathrm{W} \times \mathrm{H}$ are input dimensions and s is the subsampling factor due to pooling. Now the number of outputs of the last layer can be redimensioned to whatever is the number of classes in the specific application and the network is ready to be fine-tuned for the semantic segmentation task.

Convolutional application of ImageNet architectures typically results in considerable downsampling of the output activations with respect to the input image. Some researches have countered this effect with trained upsampling [22] which may be reinforced by taking into account switches from the strided pooling layers [16,25]. Other ways to achieve the same goal include interleaved pooling [27] and dilated convolutions [22,31]. These approaches typically improve the baseline performance in the vicinity of the object borders due to more accurate upsampling of the semantic maps. We note that the system presented in our experiments does not feature any of these techniques, however it still succeeds to deliver competitive performance.

As emphasized in the introduction, presenting a pixel-level classifier with a variety of local image representations is likely to favor the semantic segmentation performance. Previous researchers have devised two convolutional approaches to meet this idea: the skip architecture and the shared multi-class architecture. The shared multi-class architectures concatenate activations obtained by applying the common pixel-level classifier at multiple levels of the image pyramid [4,11,20,21,29]. The skip architectures concatenate activations from different levels of the deep convolutional hierarchy [22,24]. Both architectures have their merits. The shared multi-scale architecture is able to associate the evaluated window with the training dataset at unseen scales. The skip architecture allows to model the object appearance [30] and the surrounding context at different levels of detail, which may be especially appropriate for small objects. Thus it appears that best results might be obtained by a combined approach which we call a multi-scale skip architecture. The combined approach concatenates pixel representations taken at different image scales and at different levels of the deep network. We note that this idea does not appear to have been addressed in the previous work.

Despite extremely large receptive fields, pixel-level classification may still fail to establish consistent activations in all cases. Most problems of this kind concern smooth parts of very large objects: for example, a pixel in the middle of a tram may get classified as a bus. A common approach to such problems is to require pairwise agreement between pixel-level labels, which leads to a global optimization across the entire image. This requirement is often formulated as MAP inference in conditional random fields (CRF) with unary, pairwise [3,18,21] and higher-order potentials [1]. Early methods allowed binary potentials exclusively between neighboring pixels, however, this requirement has later been relaxed by defining binary potentials as linear combinations of Gaussian kernels. In this case, the message passing step in approximate mean field inference can be expressed as a convolution with a truncated Gaussian kernel in the feature space [3,18]. Recent state of the art approaches couple the CRF training and inference in custom convolutional [21] and recurrent [1] deep neural networks. We note that our present experiments feature a separately trained CRF with Gaussian potentials while our future work shall include joint CRF training [21].

We now review the details of the previous research which is most closely related to our contributions. Banica et al. [2] exploit the depth sensed by RGBD sensors to improve the quality of region proposals and the subsequent region-level classification on indoor datasets. Chen et al. [4] propose a scale attention mechanism to combine classification scores at different scales. This results in soft pooling of the classification scores across different classes. Ladicky and Shi [20] propose to train binary pixel-level classifiers which detect semantic labels at some canonical depth by exploiting the depth groundtruth obtained by LIDAR. Their inference jointly predicts semantic segmentation and depth by processing multiple levels of the monocular image pyramid. Unlike all previous approaches, our technique achieves efficient classification and training due to scale-invariant image representation recovered by exploiting reconstructed depth. Unlike [4], we perform hard scale selection at the representation level. Unlike [20], we exploit the reconstructed instead of the groundtruth depth.

## 3 Fully Convolutional Architecture with Scale Selection

We integrate the proposed technique into an end-to-end trained fully convolutional architecture illustrated in Fig. 1. The proposed architecture independently feeds images from an N-level image pyramid into the shared feature extraction network. Features from the lower levels of the pyramid are upsampled to match the resolution of features from the original image. We forward the recovered multi-scale representation to the pixel-wise scale selection multiplexer. The multiplexer is responsible for establishing a scale-invariant image representation in accordance with the reconstructed depth at the particular pixel. The back-end classifier scores the scale-invariant features with a multi-class classification model. The resulting score maps are finally converted into the per-pixel distribution over classes by a conventional softmax layer.

### 3.1 Input Pyramid and Depth Reconstruction

The left image of the input stereo pair is iteratively subsampled to produce a multi-scale pyramid representation. The first pyramid level contains the left input image. Each successive level is obtained by subsampling its predecessor with the factor $\alpha$. If the original image resolution is $\mathrm{W} \times \mathrm{H}$, the resolution of the l-th level is $\mathrm{W}/\alpha^l \times \mathrm{H}/\alpha^l$, $l \in [0..\mathrm{N}-1]$ ($\alpha = 1.3$, $\mathrm{N} = 8$). We reconstruct the depth by employing a deep correspondence metric [32] and the SGM [13] smoothness prior. The resolution of the disparity image is $\mathrm{W} \times \mathrm{H}$.

### 3.2 Single Scale Feature Extraction

Our single scale feature extraction architecture is based on the feature extraction front-end of the 16-level deep VGG-D network [29] up to the `relu5_3` layer (13 weight layers total). In order to improve the training, we introduce a batch normalization layer [14] before each non-linearity in the 5th group: `relu5_1`,

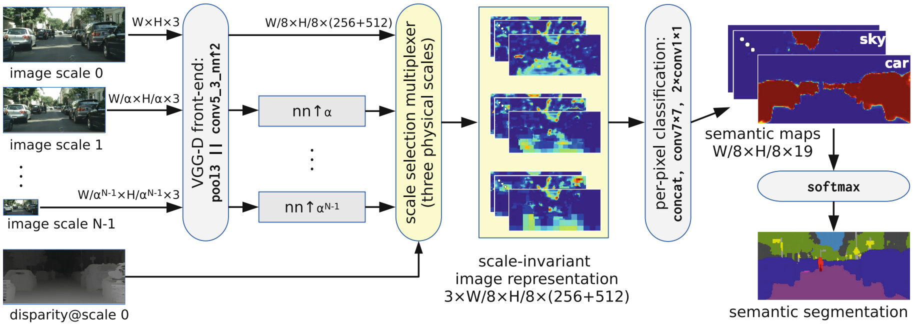

**Fig. 1.** A convolutional architecture with scale-invariant representation.

relu5_2 and relu5_3. This modification helps the fine-tuning by increasing the flow of the gradients during backprop. Subsequently we perform a $2 \times 2$ nearest neighbor upsampling of relu5_3 features and concatenate them with pool3 features in the spirit of skip architectures [22,24] (adding pool4 features did not result in significant benefits). In comparison with relu5_3, the representation from pool3 has a $2 \times 2$ higher resolution and a smaller receptive field ($40 \times 40$ vs $185 \times 185$). We hypothesize that this saves some network capacity because it relieves the network from propagating small objects through all 13 convolutional and 4 pooling layers.

The described feature extraction network is independently applied at all levels of the pyramid, in the spirit of the shared multi-class architectures [4,11,20,21,29]. Subsequently, we upsample the representations of pyramid levels 1 to N−1 in order to revert the effects of subsampling and to restore a common resolution across the representations at all scales. We perform the upsampling by a nearest neighbor algorithm in order to preserve the sparsity of the features. After upsampling, all N feature tensors have the resolution $W/8 \times H/8 \times (512 + 256)$. The $8 \times 8$ subsampling is due to three pooling levels with stride 2. Features from relu5_3 have 512 dimensions while features from pool3 have 256 dimensions.

The described procedure produces a multi-scale convolutional representation of the input image. A straight-forward approach to exploit this representation would be to concatenate the features at all N scales. However, that would imply a huge computational complexity which would make training and inference infeasible. One could also perform such procedure at some subset of scales. However, that would require a costly validation to choose the subset, while providing less information to the back-end classification network. Consequently, we proceed towards achieving scale-invariance as the main contribution of our work.

### 3.3 Scale Selection Multiplexer

The responsibility of the scale selection multiplexer is to represent each image pixel with scale-invariant convolutional features extracted at exactly $M = 3$ out of

N levels of the pyramid. The scale invariance is achieved by choosing the pyramid levels in which the apparent size of the reference metric scales are closest to the receptive field of our features.

In order to explain the details, we first establish the notation. We denote the image pixels as $p_i$, the corresponding disparities as $d_i$, the stereo baseline as $b$, and the reconstructed depths as $Z_i$. We then denote the width of the receptive field for our largest features (`conv5_3`) as $w_{rf} = 185$, the metric width of its back-projection at distance $Z_i$ as $W_i$ and the three reference metric scales in meters as $W_R = \{1, 4, 7\}$. Finally, we define $s_{mi}$ as the ratio between the m-th reference metric scale $W_{Rm}$ and the back projection $W_i$ of the receptive field:

$$s_{mi} = \frac{W_{Rm}}{W_i} = \frac{W_{Rm}}{\frac{b}{d_i} w_{rf}} = \frac{d_i \cdot W_{Rm}}{b \cdot w_{rf}} \; . \tag{1}$$

The ratio $s_{mi}$ represents the exact image scaling factor by which we should downsample the original image to attain the reference scale $m$ at pixel $i$. Now we are able to choose the representation from the pyramid level $l_{mi}$ which has a downsampling factor closest to the true factor $s_{mi}$:

$$\hat{l}_{mi} = \operatorname*{argmin}_{l} \left| \alpha^{l} - s_{mi} \right| , \quad l \in \{0, 1, ..., N-1\} \; . \tag{2}$$

The multiplexer determines the routing information at pixel $p_i$, by mapping each of the $M$ reference scales to the corresponding pyramid level $l_{mi}$. We illustrate the recovered $l_{mi}$ in Fig. 2 by color coding the computed pyramid levels at three reference metric scales. Note that in case when $s_{mi} < 1$ we simply always choose the first pyramid level ($l = 0$). We have not experimented with upsampled levels of the pyramid mostly because of memory limitations. The output of the multiplexer is a scale-invariant image representation which is stored in $M$ feature tensors of the dimension W/8 × H/8 × (512 + 256).

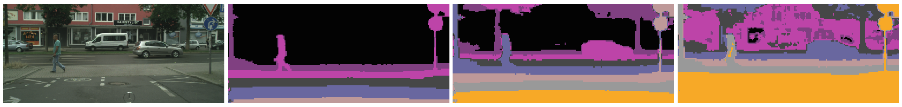

**Fig. 2.** Visualization of the scale selection switches. From left to right: original image, switches for the three reference metric scales of 1 m, 4 m and 7 m.

### 3.4 Back-End Classifier

The scale-invariant feature tensors are concatenated and passed on to the classification subnetwork which consists of one 7 × 7 and two 1 × 1 convolution+ReLU layers. The former two layers have 1024 maps and batch normalization before non-linearities. The last 1 × 1 convolutional layer is configured in a way that the number of feature maps corresponds to the number of classes. The resulting class scores are passed to the pixel-wise softmax layer to obtain the distribution across classes for each pixel.

## 4 Experiments

We evaluate our method on two different semantic segmentation datasets containing outdoor traffic scenes: Cityscapes [6] and KITTI [12]. The Cityscapes dataset [6] has 19 classes recorded in 50 cities during several months. The dataset features good and medium weather conditions, large number of dynamic objects, varying scene layout and varying background. It consists of 5000 images with fine annotations and 20000 images with coarse annotations (we use only the fine annotations). The resolution of the images is $2048 \times 1024$. The dataset includes the stereo views which we use to reconstruct the depth.

The KITTI dataset [12] provides a large collection of $1241 \times 376$ traffic videos with LIDAR reconstruction groundtruth. Unfortunately, there are no official semantic segmentation annotations for this dataset. However, a collection of 150 images annotated with 11 object classes has been published [26]. We expand that work by annotating the same 11 classes in another 299 images from the same dataset, as well as by fixing some inconsistent annotations in the original dataset. The combined dataset with 399 training and 46 test images is freely available for academic research[1].

We train our networks using Adam SGD [17] and batch normalization [14] without learnable parameters. Due to memory limitations we only have one image in a batch. The input images are zero-centered and normalized. We initialize the learning rate to $10^{-5}$, decrease it to $0.5 \cdot 10^{-5}$ after 2nd epoch and again to $10^{-6}$ after 10th epoch. Before each epoch, the training set is shuffled to eliminate the bias. The first 13 convolutional layers are initialized from VGG-D [29] pretrained on ImageNet and fine tuned during training. All other layers are randomly initialized. In all experiments we train our networks for 15 epochs on Cityscapes dataset and 30 epochs on KITTI dataset. We use the softmax cross-entropy loss which is summed up over all the pixels in a batch. In both datasets the frequency of pixel labels is highly unevenly distributed. We therefore perform class balancing by weighting each pixel loss with the true class weight factor $w_c$. This factor can be determined from class frequencies in the training set as follows: $w_c = \min(10^3, p(c)^{-1})$, where $p(c)$ is the frequency of pixels from the class $c$ in the batch. In all experiments the three reference metric scales were set to equidistant values $W_R = \{1, 4, 7\}$ which were not cross-validated but had a good coverage over the input pixels (cf. Fig. 2). A Torch implementation of this procedure is freely available for academic research[2].

In order to alleviate the downsampling effects and improve consistency, we postprocess the semantic map with the fully-connected CRF [18]. The negative logarithms of probability distributions across classes are used as unary potentials, while the pairwise potentials are based on a linear combination of two Gaussian kernels [18]. We fix the number of mean field iterations to 10. The smoothness kernel parameters are fixed at $w^{(2)} = 3, \theta_\gamma = 3$, while a coarse grid

[1] http://multiclod.zemris.fer.hr/kitti_semseg_unizg.shtml.

[2] https://github.com/ivankreso/scale-invariant-cnn.

**Table 1.** Individual class results on the Cityscapes validation and test sets (IoU scores). Last row represents a fraction by which each class is represented in the training set.

| method | road | sidewalk | building | wall | fence | pole | traffic light | traffic sign | vegetation | terrain | sky | person | rider | car | truck | bus | train | motorcycle | bicycle | mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Validation set | | | | | | | | | | | | | | | | | | | | |
| Single5 | 93.0 | 64.6 | 80.6 | 28.7 | 37.5 | 30.4 | 38.5 | 43.6 | 81.9 | 42.3 | 81.5 | 57.2 | 37.3 | 83.4 | 36.0 | 43.9 | 37.1 | 38.6 | 58.8 | 53.4 |
| Single3+5 | 93.8 | 67.9 | 83.2 | 30.1 | 37.4 | 37.8 | 45.3 | 55.5 | 86.3 | 49.3 | 85.9 | 63.8 | 40.9 | 87.0 | 32.7 | 50.5 | 31.7 | 37.1 | 63.5 | 56.8 |
| FixedScales | 94.6 | 70.9 | 85.5 | 42.5 | 42.4 | 39.6 | 46.2 | 55.3 | 86.5 | 50.3 | **86.4** | 64.4 | 45.4 | 88.3 | 47.2 | 60.8 | 54.0 | 43.1 | 63.9 | 61.4 |
| ScaleInvariant | **95.1** | **72.4** | **86.6** | **45.3** | **47.3** | **42.0** | **50.6** | **56.3** | **87.2** | **52.7** | 86.1 | **69.8** | **51.1** | **89.0** | **55.3** | **70.8** | **54.0** | **46.8** | **64.1** | **64.4** |
| Test set | | | | | | | | | | | | | | | | | | | | |
| ScaleInvariant | 95.3 | 73.5 | 86.4 | 36.8 | 42.7 | 45.5 | 56.6 | 57.8 | 89.5 | 63.6 | 90.3 | 73.5 | 53.1 | **90.3** | 30.8 | 48.2 | 39.6 | 52.2 | 62.7 | 62.5 |
| ScaleInvariant+CRF | **96.3** | **76.8** | **88.8** | **40.0** | **45.4** | **50.1** | **63.3** | **69.6** | **90.6** | **67.1** | **92.2** | **77.6** | **55.9** | 90.1 | **39.2** | **51.3** | **44.4** | **54.4** | **66.1** | **66.3** |
| % in train | 33.5 | 6.3 | 24.0 | 0.7 | 1.0 | 1.4 | 0.2 | 0.6 | 16.9 | 1.2 | 4.0 | 1.4 | 0.2 | 7.6 | 0.3 | 0.3 | 0.3 | 0.1 | 0.5 | |

search on 200 Cityscapes images is performed to optimize $w^{(1)} \in \{5, 10\}, \theta_\alpha \in \{50, 60, 70, 80, 90\}, \theta_\beta \in \{3, 4, 5, 6, 7, 8, 9\}$.

The segmentation performance is measured by the intersection-over-union (IoU) score [10] and the pixel accuracy [22]. We first evaluate the performance of two single scale networks which are obtained by eliminating the scale multiplexer layer and applying our network to full resolution images (cf. Fig. 1). This network is referred to as Single3+5. Furthermore, we also have the Single5 network which is just Single3+5 without the representation from `pool3`. The label ScaleInvariant shall refer to the full architecture visualized in Fig. 1. The label FixedScales refers to the similar architecture with fixed multi-scale representation obtained by concatenating the pyramid levels 0, 3 and 7.

First, we show results on the Cityscapes dataset. We downsample original images and train on smaller resolution (1504 × 672) due to memory limitations. Table 1 shows the results on the validation and test sets. We can observe that our scale invariant network improves over the single scale approach across all classes in Table 1. We can likewise notice that the concatenation from pool3 is important as Single3+5 produces better results then Single3 and that improvement is larger for smaller classes like poles, traffic signs and traffic lights. That supports our hypothesis that the representation from `pool3` helps to better handle smaller objects. Furthermore, the scale invariant network achieves significant improvement over multi-scale network with fixed image scale levels (FixedScales). This agrees with our hypothesis that the proposed scale selection approach should help the network to learn a better representation. The table also shows results on the test set (our online submission is entitled Scale invariant CNN + CRF).

The last row in Table 1 represents a proportion by which each class is represented in the training set. We can notice that the greatest contribution of our approach is achieved for classes which represent smaller objects or objects that we see less often like buses, trains, trucks, walls etc. This effect is illustrated in Fig. 3 where we plot the improvement of the IoU metric with respect to the training set proportion for each class. Likewise we achieve improvement in pixel accuracy from 90.1 % (Single3+5) to 91.9 % (ScaleInvariant).

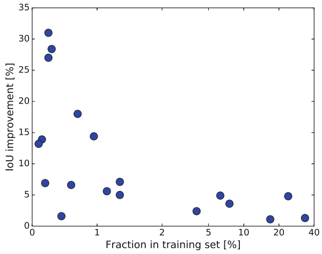

**Fig. 3.** Improvement of the IoU metric between Single3+5 and ScaleInvariant architecture with respect to the proportion inside training set for each class.

Figure 4 shows examples where the scale-invariant network produces better results. The improvement for big objects is clearly substantial. We often observe that the scale-invariant network can differentiate between road and sidewalk and person and rider, especially when they assume a rare appearance as in the last row with cobbled road which is easily mistaken for sidewalk.

Table 2 shows results on the KITTI test set. We notice a significant improvement in mean IoU class metric, which is, however, smaller than on Cityscapes. The main reason is that KITTI has much less smaller classes (only pole, sign and cyclist). Furthermore, it is a much smaller dataset which explains why the

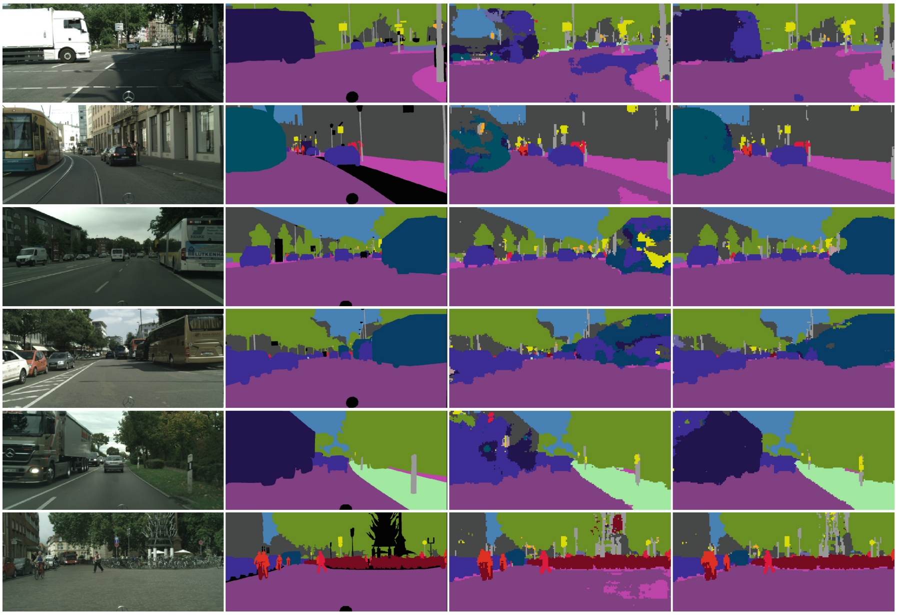

**Fig. 4.** Examples where scale invariance helps the most. From left to right: input, groundtruth, baseline segmentation (Single3+5), scale invariant segmentation.

**Table 2.** Individual class results on the KITTI test set (IoU scores).

| Method | Sky | Building | Road | Sidewalk | Fence | Vegetation | Pole | Car | Sign | Pedestrian | Cyclist | Mean |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Single3+5 | 82.1 | 82.1 | 89.5 | 71.8 | 37.1 | **80.7** | 27.4 | 79.7 | 32.6 | 41.9 | **14.3** | 58.1 |
| ScaleInvariant | **84.1** | **83.9** | **91.4** | **73.1** | **43.2** | 79.2 | **33.6** | **82.0** | **44.7** | **57.3** | 12.8 | **62.3** |

performance is so low on classes like cyclist and pole. Here we again report an improvement in pixel accuracy from 87.63 (Single3+5) to 88.57 (ScaleInvariant).

## 5 Conclusion

We have presented a novel technique for improving semantic segmentation performance. We use the reconstructed depth as a guide to produce a scale-invariant representation in which the appearance is decoupled from the scale. This precludes the necessity to recognize objects at all possible scales and allows for an efficient use of the classifier capacity and the training data. This trait is especially important for navigation datasets which contain objects at a great variety of scales and do not exhibit the photographer bias.

We have integrated the proposed technique into an end-to-end trainable fully convolutional architecture which extracts features by a multi-scale skip network. The extracted features are fed to the novel multiplexing layer which carries out dense scale selection at the pixel level and produces a scale-invariant representation which is scored by the back-end classification network.

We have performed experiments on the novel Cityscapes dataset. Our results are very close to the state-of-the-art, despite the fact that we have trained our network on reduced resolution. We also report experiments on the KITTI dataset where we have densely annotated 299 new images, improved 146 already available annotations and release the union of the two datasets to the community. The proposed scale selection approach has consistently contributed substantial increases in segmentation performance. The results show that deep neural networks are extremely powerful classification models, however, they are still unable to learn geometric transformation better than humans.

**Acknowledgement.** This work has been fully supported by Croatian Science Foundation under the project I-2433-2014. The Titan X used for this research was donated by the NVIDIA Corporation.

## References

1. Arnab, A., Jayasumana, S., Zheng, S., Torr, P.H.S.: Higher order potentials in end-to-end trainable conditional random fields. CoRR abs/1511.08119 (2015)
2. Banica, D., Sminchisescu, C.: Second-order constrained parametric proposals and sequential search-based structured prediction for semantic segmentation in RGB-D images. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, pp. 3517–3526, 7–12 June 2015
3. Chen, L., Papandreou, G., Kokkinos, I., Murphy, K., Yuille, A.L.: Semantic image segmentation with deep convolutional nets and fully connected crfs. In: International Conference on Learning Representations, ICLR 2015, San Diego, California (2014)
4. Chen, L., Yang, Y., Wang, J., Xu, W., Yuille, A.L.: Attention to scale: scale-aware semantic image segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Las Vegas, Nevada (2016) (to appear)
5. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A., Ma, H., Fidler, S., Urtasun, R.: 3d object proposals for accurate object class detection. In: NIPS (2015)
6. Cordts, M., Omran, M., Ramos, S., Scharwächter, T., Enzweiler, M., Benenson, R., Franke, U., Roth, S., Schiele, B.: The cityscapes dataset. In: CVPR Workshop on the Future of Datasets in Vision (2015)
7. Divvala, S.K., Hoiem, D., Hays, J., Efros, A.A., Hebert, M.: An empirical study of context in object detection. In: 2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2009), Miami, Florida, USA, pp. 1271–1278, 20–25 June 2009
8. Eigen, D., Fergus, R.: Predicting depth, surface normals and semantic labels with a common multi-scale convolutional architecture. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, pp. 2650–2658, 7–13 December 2015
9. Everingham, M., Eslami, S.M.A., Gool, L.V., Williams, C.K.I., Winn, J.M., Zisserman, A.: The pascal visual object classes challenge: a retrospective. Int. J. Comput. Vis. **111**(1), 98–136 (2015)
10. Everingham, M., Gool, L., Williams, C.K., Winn, J., Zisserman, A.: The pascal visual object classes (voc) challenge. Int. J. Comput, Vis. **88**(2), 303–338 (2010)
11. Farabet, C., Couprie, C., Najman, L., LeCun, Y.: Learning hierarchical features for scene labeling. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1915–1929 (2013)
12. Geiger, A., Lenz, P., Stiller, C., Urtasun, R.: Vision meets robotics: the kitti dataset. Int. J. Robot. Res. (IJRR) (2013)
13. Hirschmüller, H.: Stereo vision in structured environments by consistent semi-global matching. In: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2006), New York, NY, USA, pp. 2386–2393, 17–22 June 2006
14. Ioffe, S., Szegedy, C.: Batch normalization: Accelerating deep network training by reducing internal covariate shift. In: Proceedings of the International Conference on Machine Learning, ICML 2015, Lille, France, pp. 448–456, 6–11 July 2015
15. Jia, Y., Shelhamer, E., Donahue, J., Karayev, S., Long, J., Girshick, R.B., Guadarrama, S., Darrell, T.: Caffe: Convolutional architecture for fast feature embedding. In: Proceedings of the ACM International Conference on Multimedia, MM 2014, Orlando, FL, USA, 03–07 November 2014, pp. 675–678 (2014)
16. Kendall, A., Badrinarayanan, V., Cipolla, R.: Bayesian segnet: model uncertainty in deep convolutional encoder-decoder architectures for scene understanding. CoRR abs/1511.02680 (2015)

17. Kingma, D.P., Ba, J.: Adam: a method for stochastic optimization. CoRR abs/1412.6980 (2014)
18. Krähenbühl, P., Koltun, V.: Efficient inference in fully connected crfs with gaussian edge potentials. In: Advances in Neural Information Processing Systems 24: 25th Annual Conference on Neural Information Processing Systems 2011, Proceedings of a meeting held 12–14, Granada, Spain, pp. 109–117, December 2011
19. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Annual Conference on Neural Information Processing Systems, Lake Tahoe, Nevada, United States, pp. 1106–1114 (2012)
20. Ladicky, L., Shi, J., Pollefeys, M.: Pulling things out of perspective. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2014, Columbus, OH, USA, 23–28 June 2014, pp. 89–96 (2014)
21. Lin, G., Shen, C., van dan Hengel, A., Reid, I.: Efficient piecewise training of deep structured models for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR, Las Vegas, Nevada (2016) (to appear)
22. Long, J., Shelhamer, E., Darrell, T.: Fully convolutional networks for semantic segmentation. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015, pp. 3431–3440 (2015)
23. Martinovic, A., Knopp, J., Riemenschneider, H., Gool, L.V.: 3d all the way: semantic segmentation of urban scenes from start to end in 3d. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015
24. Mostajabi, M., Yadollahpour, P., Shakhnarovich, G.: Feedforward semantic segmentation with zoom-out features. In: IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2015, Boston, MA, USA, 7–12 June 2015, pp. 3376–3385 (2015)
25. Noh, H., Hong, S., Han, B.: Learning deconvolution network for semantic segmentation. In: 2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, 7–13 December 2015, pp. 1520–1528 (2015)
26. Ros, G., Ramos, S., Granados, M., Bakhtiary, A., Vázquez, D., López, A.M.: Vision-based offline-online perception paradigm for autonomous driving. In: 2015 IEEE Winter Conference on Applications of Computer Vision, WACV 2014, Waikoloa, HI, USA, 5–9 January 2015, pp. 231–238 (2015)
27. Sermanet, P., Eigen, D., Zhang, X., Mathieu, M., Fergus, R., LeCun, Y.: Overfeat: integrated recognition, localization and detection using convolutional networks. In: International Conference on Learning Representations, ICLR 2014, Banff, Canada, pp. 1–16 (2014)
28. Shotton, J., Winn, J.M., Rother, C., Criminisi, A.: Textonboost for image understanding: multi-class object recognition and segmentation by jointly modeling texture, layout, and context. Int. J. Comput. Vis. **81**(1), 2–23 (2009)
29. Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. In: International Conference on Learning Representations, ICLR 2015, San Diego, California, pp. 1–16 (2014)
30. Viola, P.A., Jones, M.J.: Robust real-time face detection. Int. J. Comput. Vis. **57**(2), 137–154 (2004)
31. Yu, F., Koltun, V.: Multi-scale context aggregation by dilated convolutions. In: International Conference on Learning Representations, ICLR 2016, San Juan, Puerto Rico, pp. 1–9 (2016)
32. Zbontar, J., LeCun, Y.: Stereo matching by training a convolutional neural network to compare image patches. CoRR abs/1510.05970 (2015)